Future Directions in Scientific Supercomputing for Computational Physics

Horst D. Simon

Director

National Energy Research Scientific Computing Center

(NERSC)

Berkeley, California, USA

CCP 2001 Aachen

September 2001

How fast things change …

Then	Now (or soon)
1969: Apollo Lunar Excursion Module 48 Kbyte ROM	2001: Rocket the Wonder Dog (toy) 256 Kbyte ROM
1985: Cray-2 supercomputer 2 Gflop/s	2001: Hello Kitty personal computer 1.8 Gflop/s
1991: Space shuttle 1 MHz onboard computer	2001: Mercedes-Benz S-500 100 MHz onboard computer
1991: SGI Indigo-2 graphics wkst. 350,000 polygons per second	2001: X-Box game console 125 million polygons per second
1996: IBM Deep Blue chess computer 200 million moves analyzed/sec	2008 (expected): Tabletop chess 1 billion moves analyzed/sec

SOURCE: Turning Powerhouses into Playthings
[from Wired, June 2001, pg. 88]

"It's hard to make predictions, especially about the future."

Yogi Berra

Overview

1) Computational Science at NERSC

2) Strategic Plan 2002 - 2006

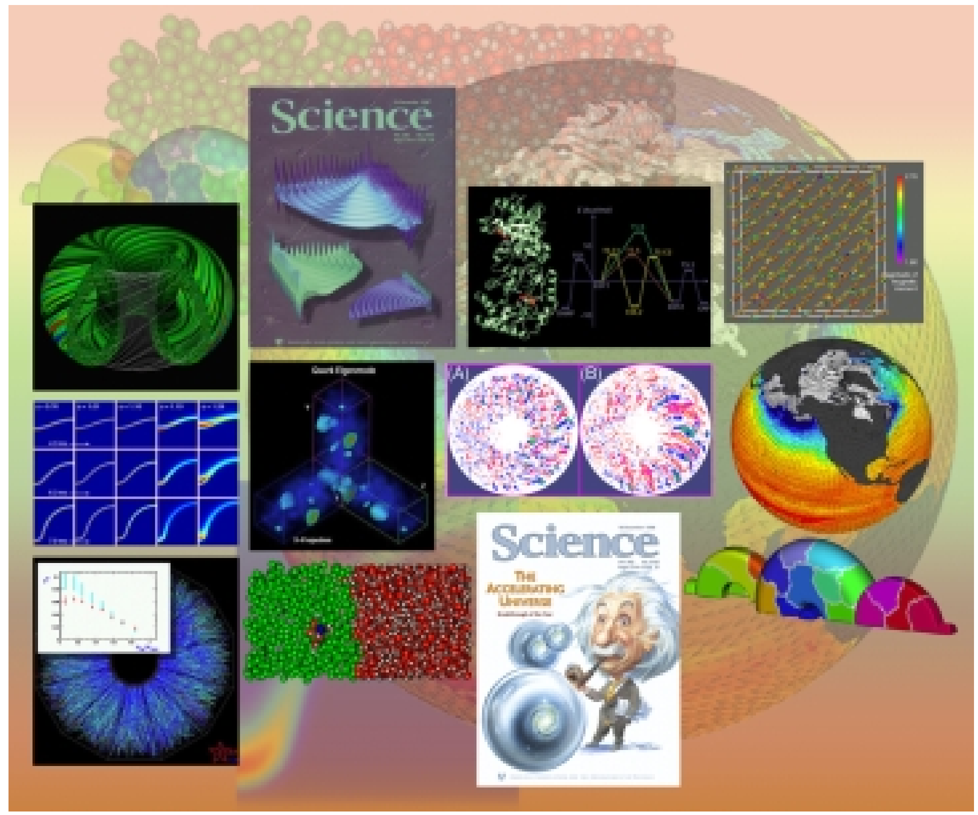
3) High Performance Computing trends in the next decade

NERSC Overview

- Located in the hills next to University of California, Berkeley campus
- close collaborations between university and NERSC in computer science and computational science

NERSC - Overview

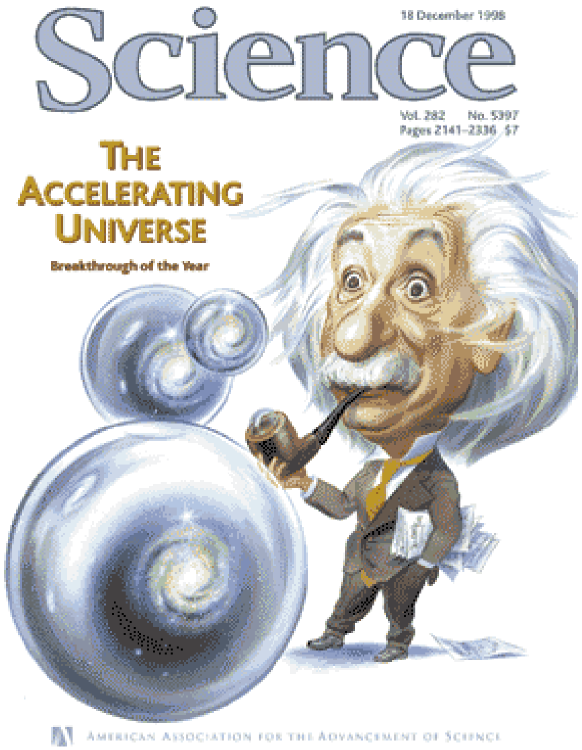
- the Department of Energy, Office of Science, supercomputer facility
- unclassified, open facility; serving >2000 users in all DOE mission relevant basic science disciplines
- 25th anniversary in 1999 (one of the oldest supercomputing centers)

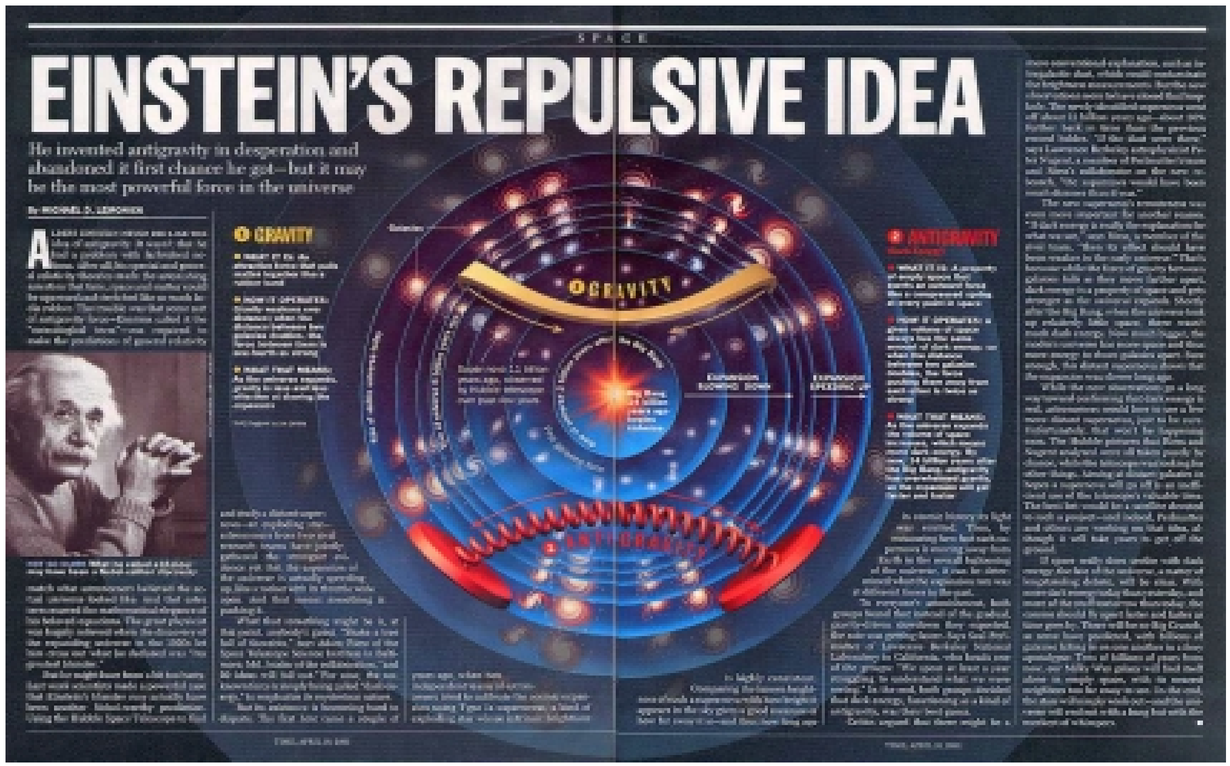


Support for Computational Cosmology

Computing for Supernova Cosmology

Over the past 3 years the observations of supernovae at high redshift has shown that the universe is currently accelerating and that over 2/3 of it is in the form of "dark energy".





SPACE

EINSTEIN'S REPULSIVE IDEA

He invented antigravity in desperation and abandoned it first chance he got—but it may be the most powerful force in the universe

Collaborations are Enabling Scientific Discoveries

- BOOMERANG Experiments – analyze cosmic microwave background radiation data to obtain a better understanding of the universe.
- The data analysis provides strong evidence that the universe is flat.
- Developed MADCAP software and provided computational capability on NERSC platforms

Nature, April 27, 2000

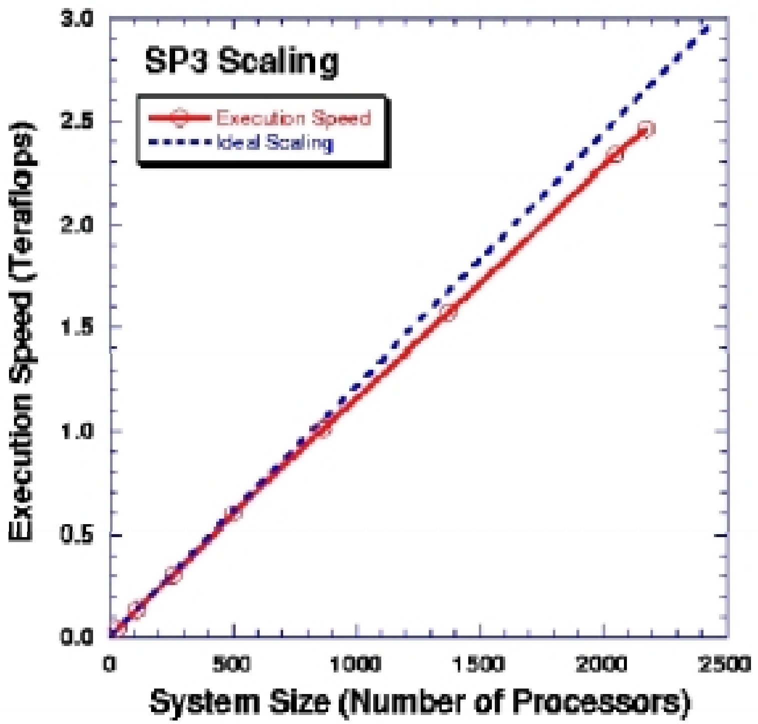
Multi-Teraflops Spin Dynamics Studies of the Magnetic Structure of FeMn/Co Interfaces

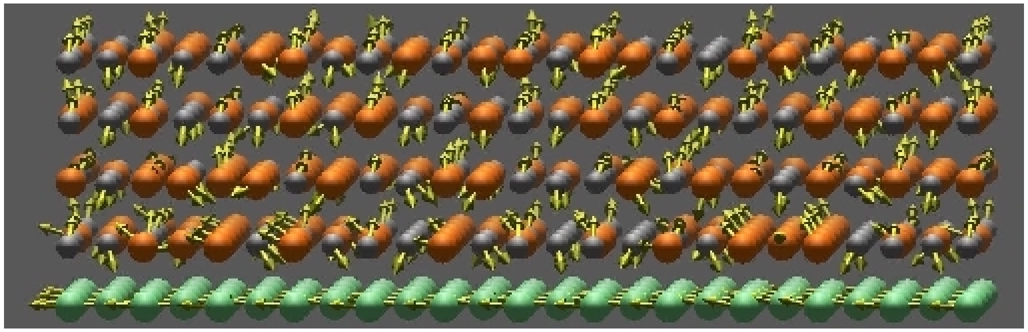
Exchange bias, which involves the use of an antiferromagnetic (AFM) layer such as FeMn to pin the orientation of the magnetic moment of approximate ferromagnetic (FM) layer such as Co, is of fundamental importance in magnetic multilayer storage and read head devices.

The full simulation used 2016 atoms ran at 2.26 Teraflops on 126 nodes.

A larger simulation of 2176 atoms of FeMn ran at 2.46 Teraflops on 136 nodes.

A. Canning et al., Proc. IEEE SC01, (to appear).





Section of an FeMn/Co (Iron Manganese/ Cobalt) interface showing the final configuration of the magnetic moments for five layers at the interface.

Shows a new magnetic structure which is different from the 3Q magnetic structure of pure FeMn.

Overview

1) Computational Science at NERSC

2) Strategic Plan 2002 - 2006

3) High Performance Computing trends in the next decade

Strategic Components of NERSC 2002 - 2006

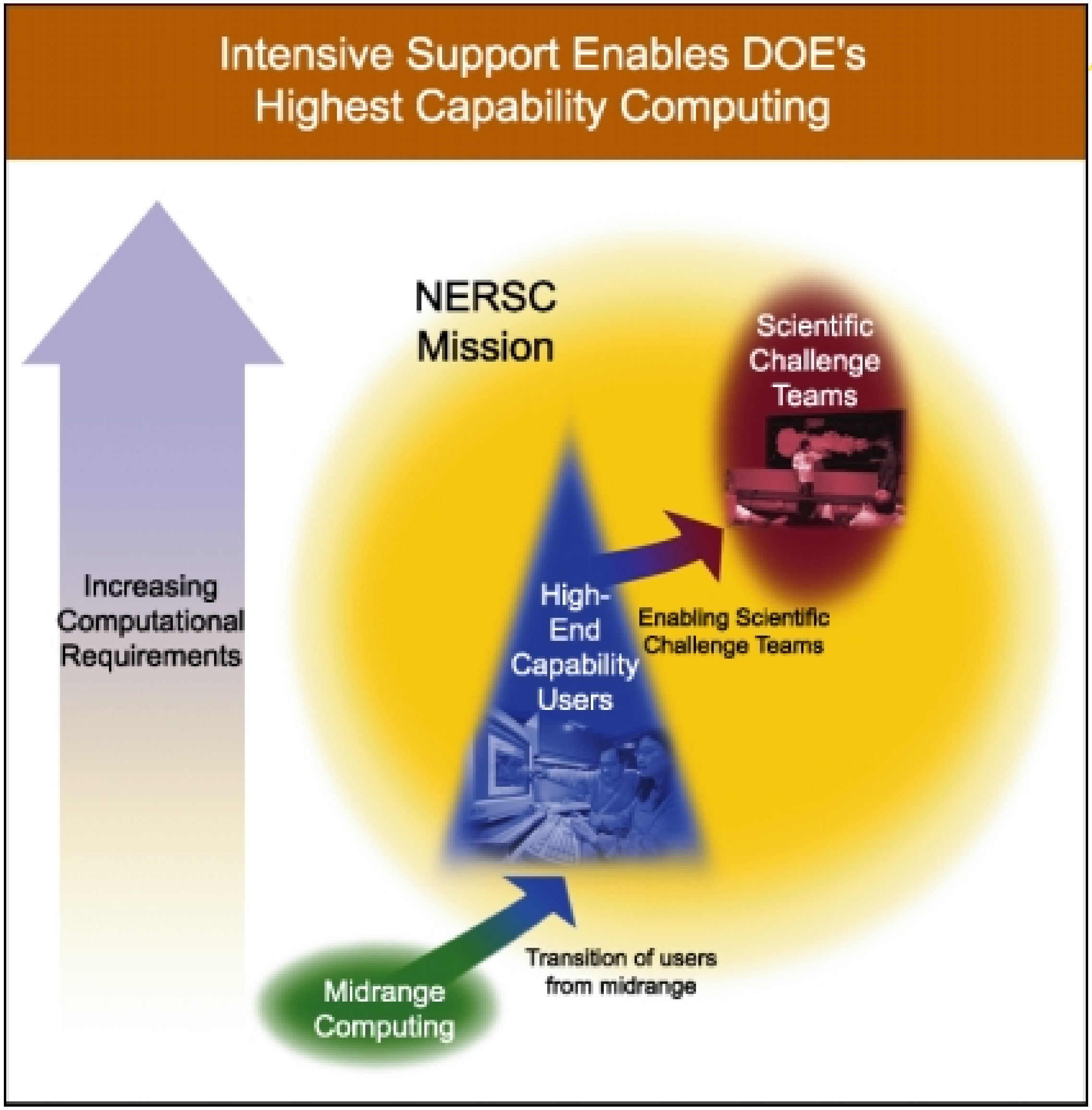
Terascale Computing at NERSC

NERSC-3

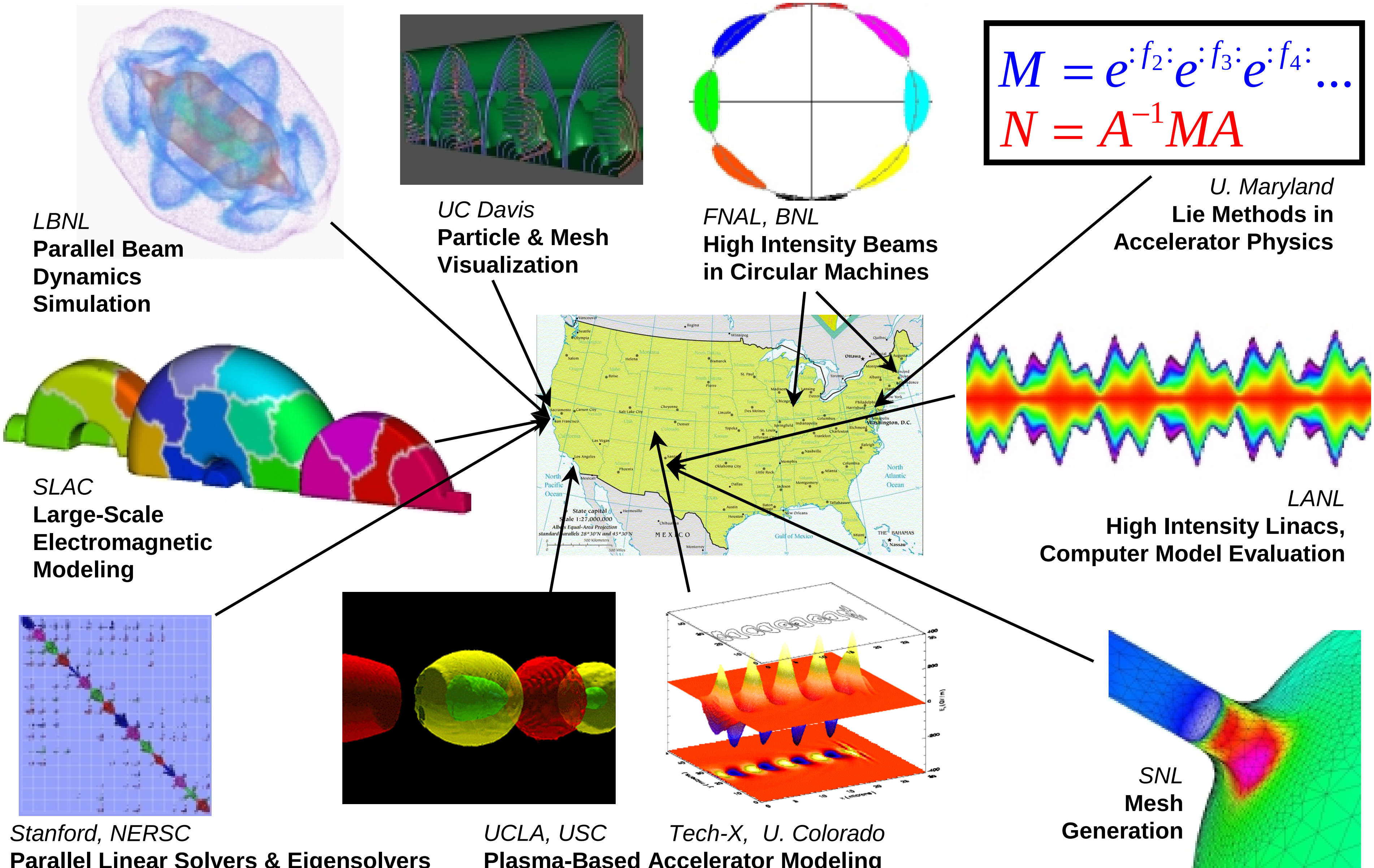
TOP500 List 11/00

RANK	MANU-FACTURER	COMPUTER	RMAX [TF/S]	INSTALLATION SITE	COUNTRY	YEAR	AREA OF INSTALLATION	# PROC
1	IBM	ASCI White SP Power3	4.93	Lawrence Livermore National Laboratory	USA	2000	Research	8192
	IBM	NERSC-3	2.526 TF/s		USA	2000	Research	2528
2	Intel	ASCI Red	2.38	Sandia National Laboratory	USA	1999	Research	9632
3	IBM	ASCI Blue Pacific SST, IBM SP 604E	2.14	Lawrence Livermore National Laboratory	USA	1999	Research	5808
4	SGI	ASCI Blue Mountain	1.61	Los Alamos National Laboratory	USA	1998	Research	6144
5	IBM	SP Power3 375Mhz	1.42	IBM/Naval Oceanographic Office (NAVOCEANO)	USA	2000	Research	1336
6	IBM	SPPower3 375Mhz	1.18	National Centers for Environmental Predicition	USA	2000	Research	1104
7	Hitachi	SR8000-F1	1.04	Leibniz Rechenzentrum, Munic	Germany	2000	Academic	112
8	IBM	SP Power3 375MHz 8way	0.93	San Diego Supercomputer Center	USA	2000	Academic	1152
9	Hitachi	SR8000-F1	0.92	High Energy Accelerator Research Organization/ KEK,	Japan	2000	Research	100
10	Cray Inc.	T3E 1200	0.89	Government	USA	1998	Classified	1084

Comprehensive Scientific Support and Enabling Science Challenge Teams



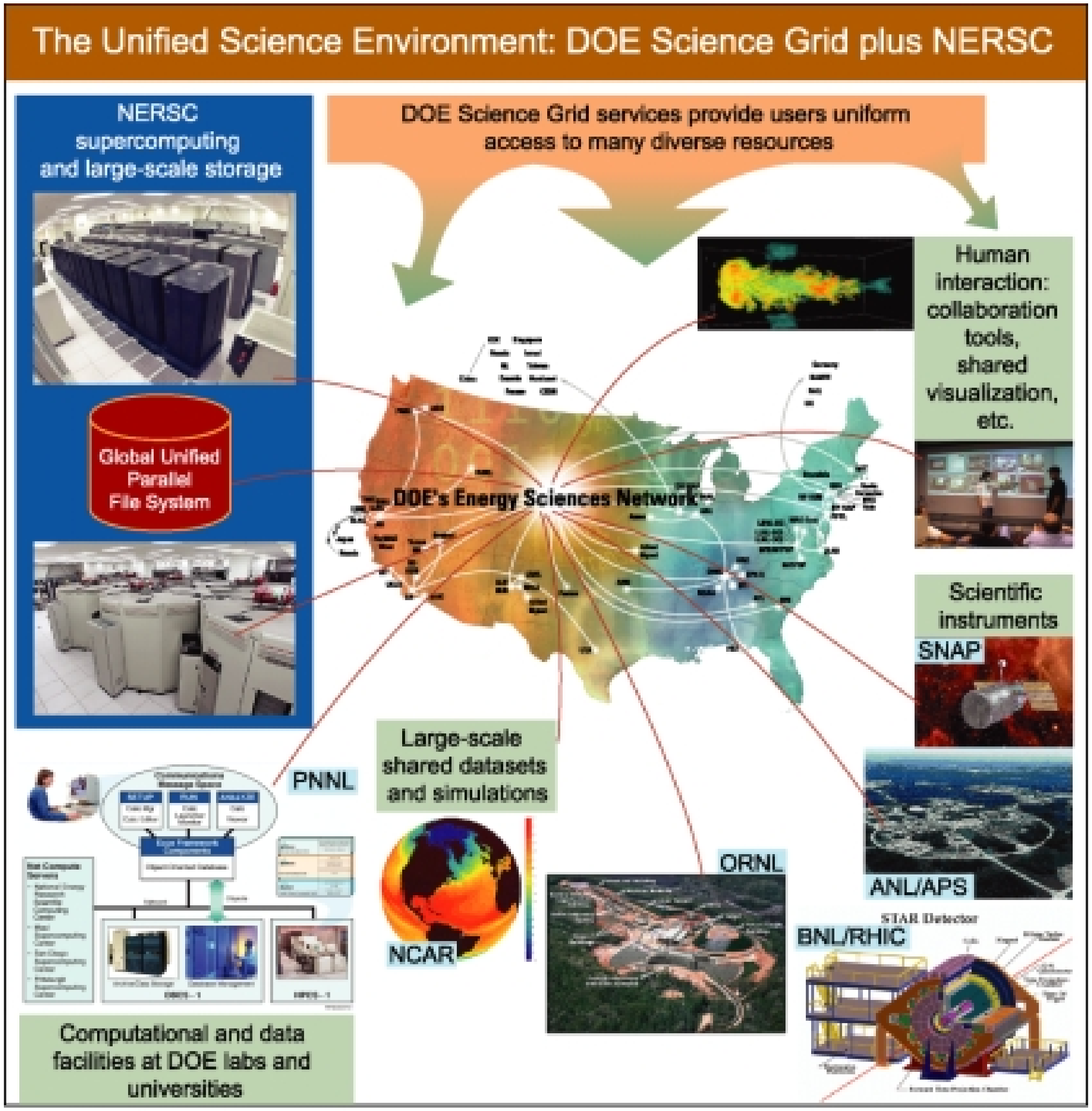
Accelerating Scientific Discovery in Accelerator Technology and Beam Physics: A SciDAC Multi-disciplinary, Multi-institutional Collaboration



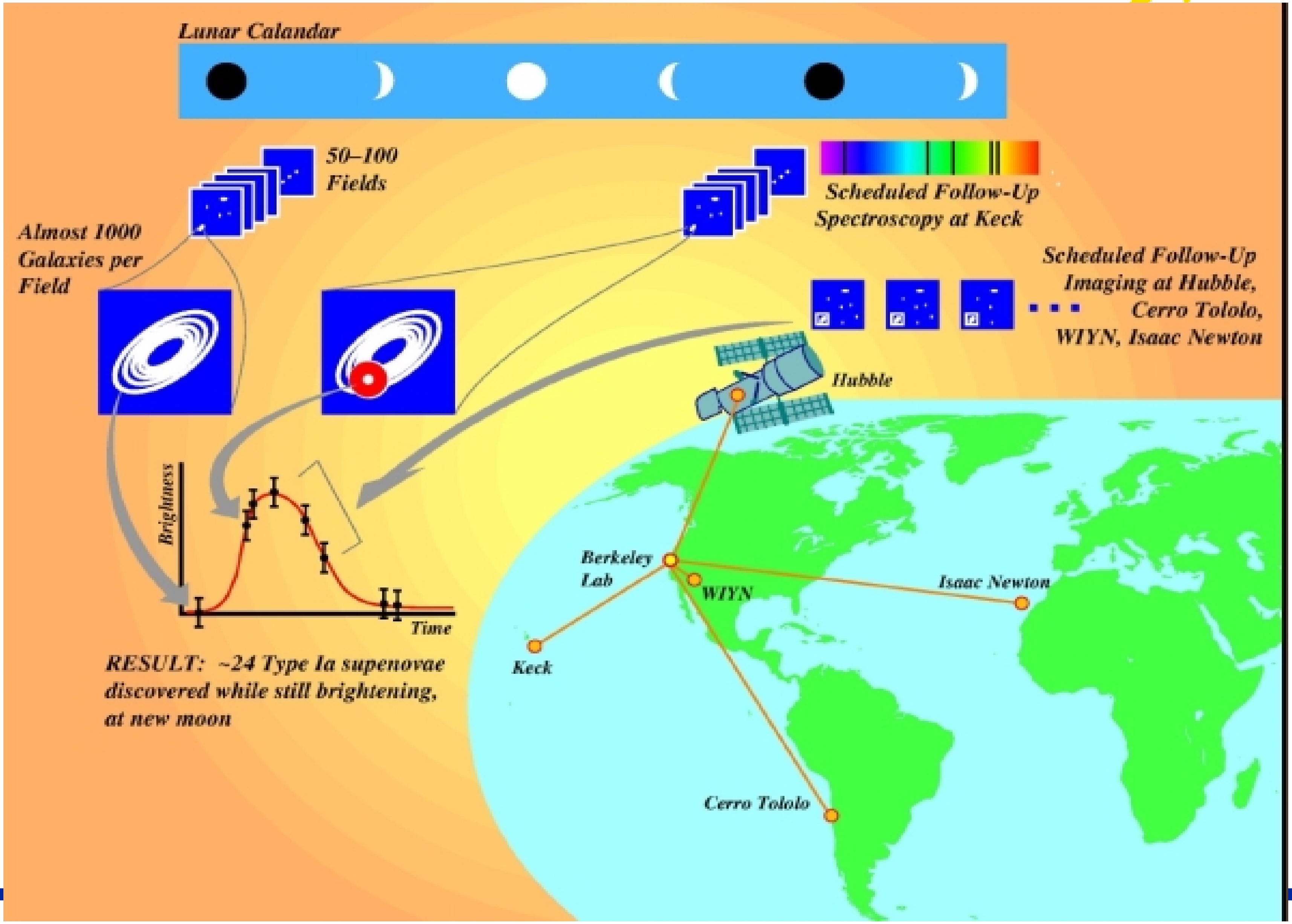
CERN and PSI are European collaborators

Courtesy Rob Ryne, LBNL

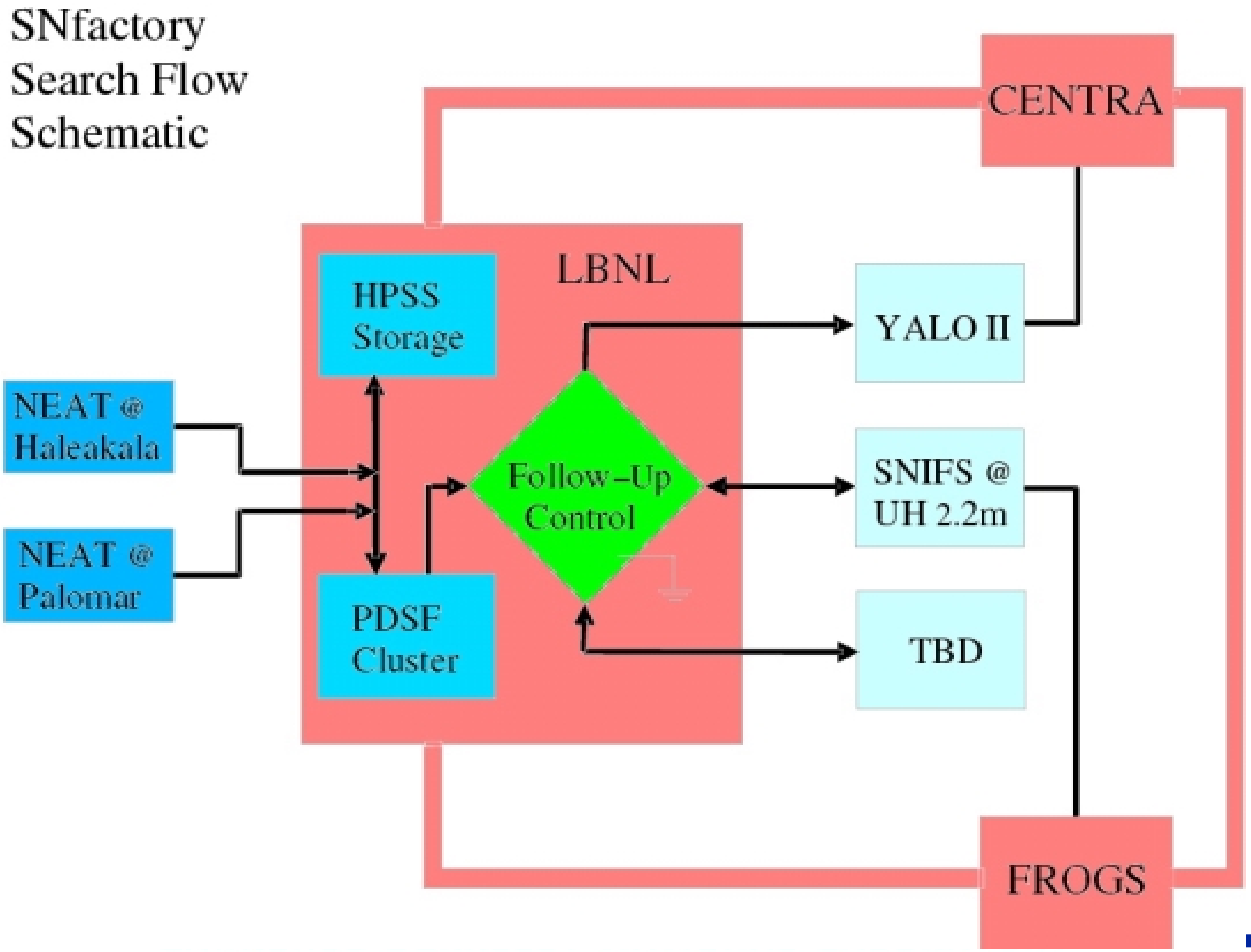
Unified Science Environment



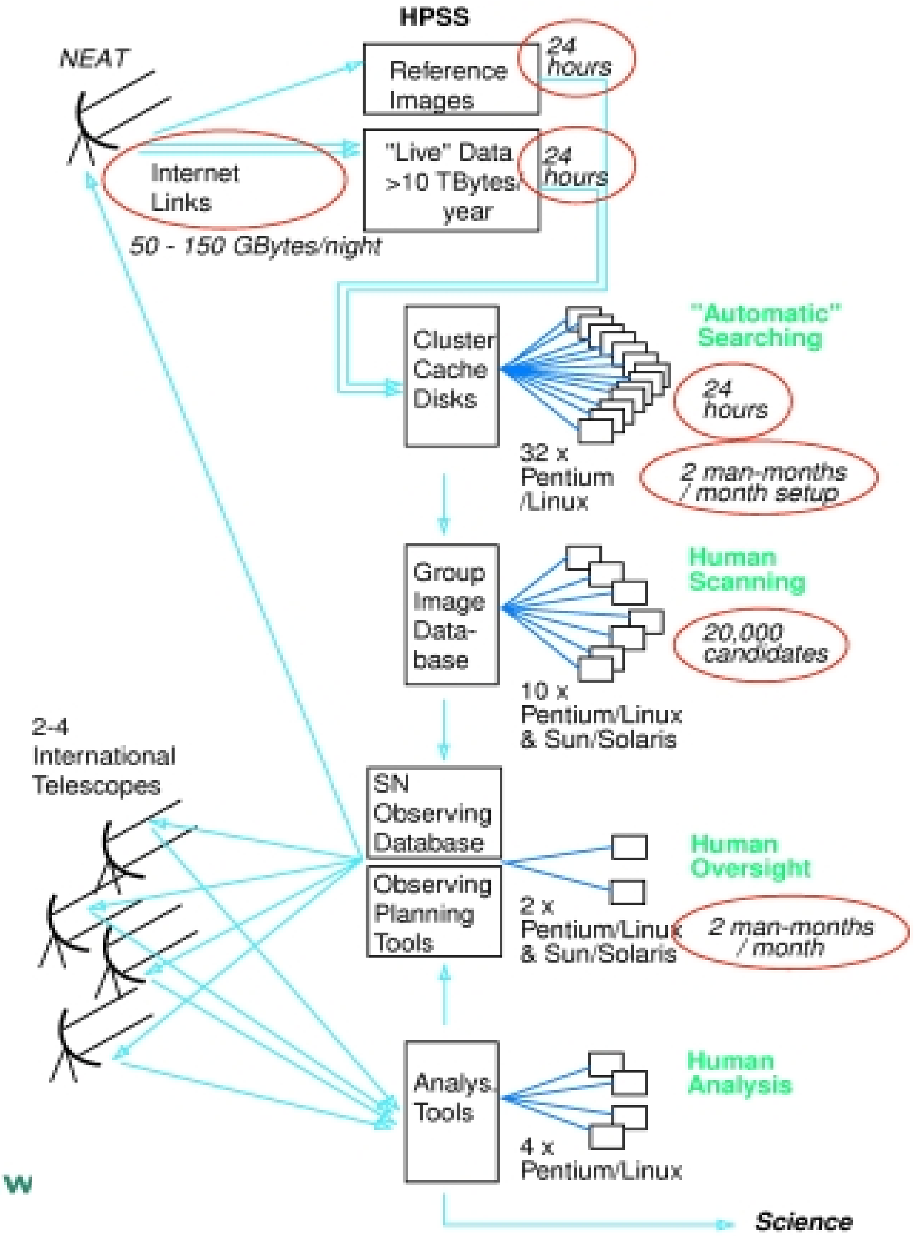
Search Strategy for Nearby Supernovae



SNfactory Search Flow



Supernova Factory



Summary on Tends in Supercomputing Centers

- Continued rapid growth of high end computational and storage resources
- Continued requirement for comprehensive scientific support
- Increasing formation of large scale, multi-institutional, multi-disciplinary collaborations
- Integration of centers into grids

Overview

1) Computational Science at NERSC

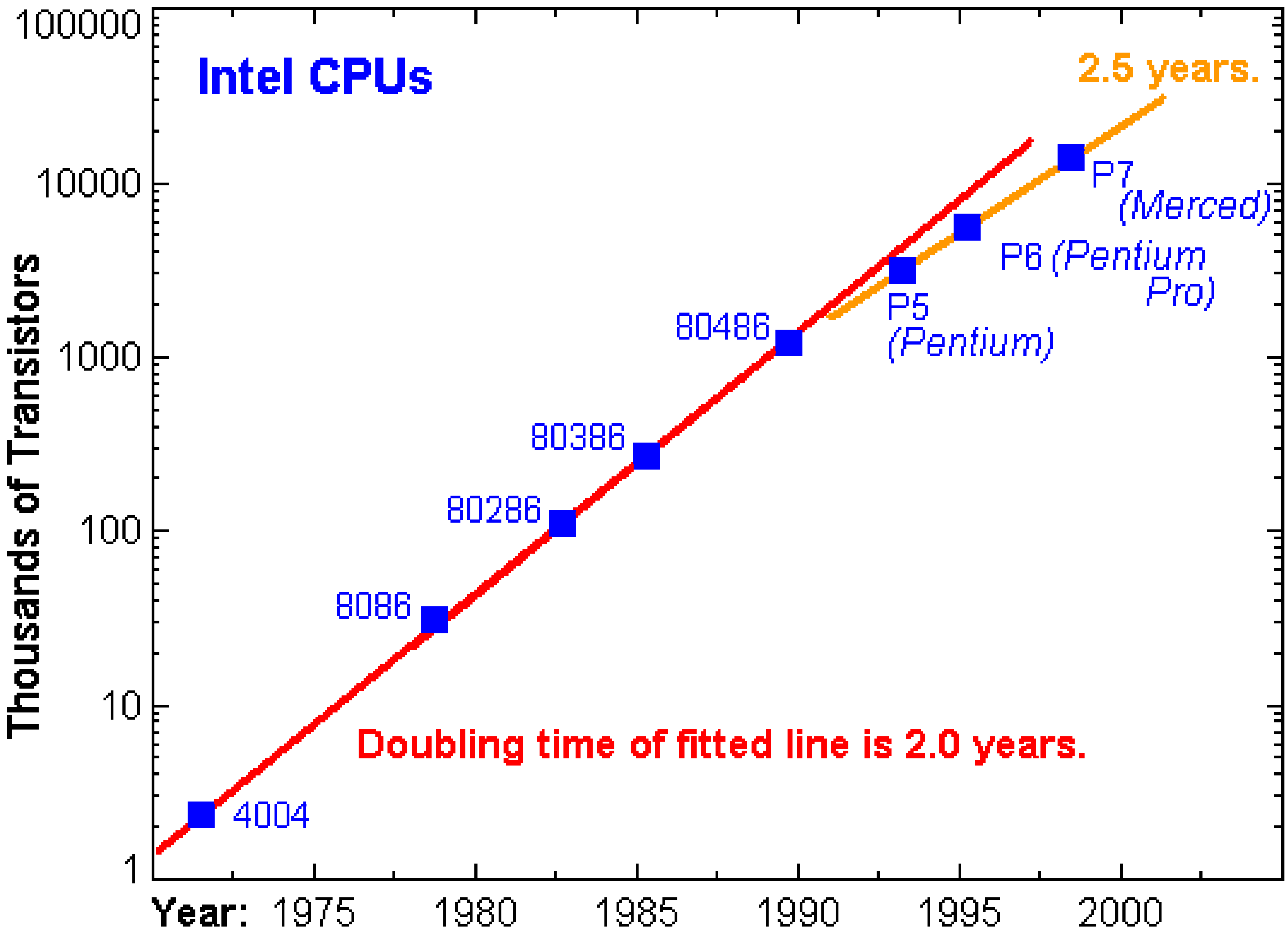
2) Strategic Plan 2002 - 2006

3) High Performance Computing trends in the next decade

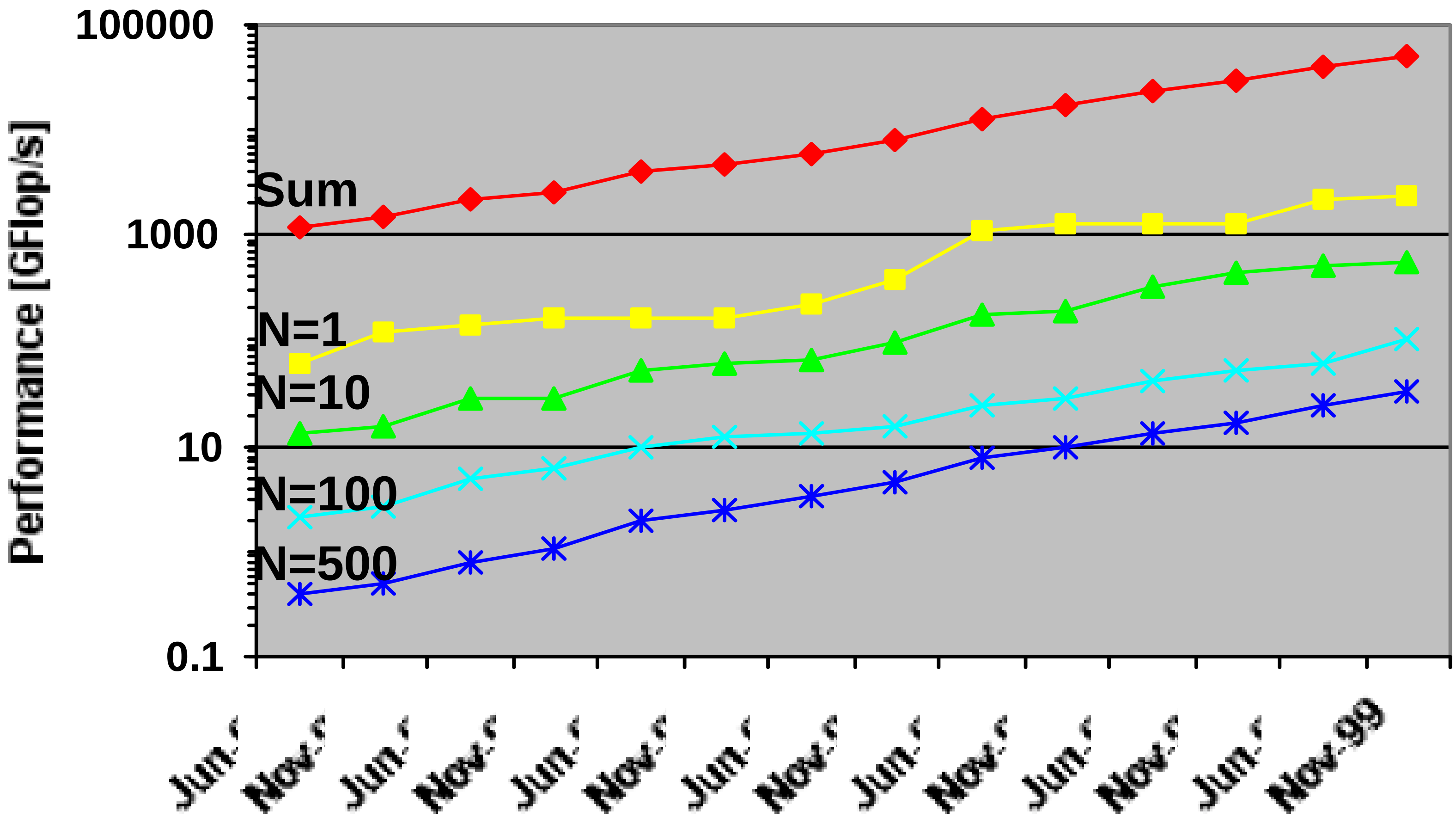
Five Computing Trends for the Next Five Years

- Continued rapid processor performance growth following Moore's law
- Open software model (Linux) will become standard
- Network bandwidth will grow at an even faster rate than Moore's Law
- Aggregation, centralization, colocation
- Commodity products everywhere

Moore's Law —
The Traditional (Linear) View



Performance Increases in the TOP500

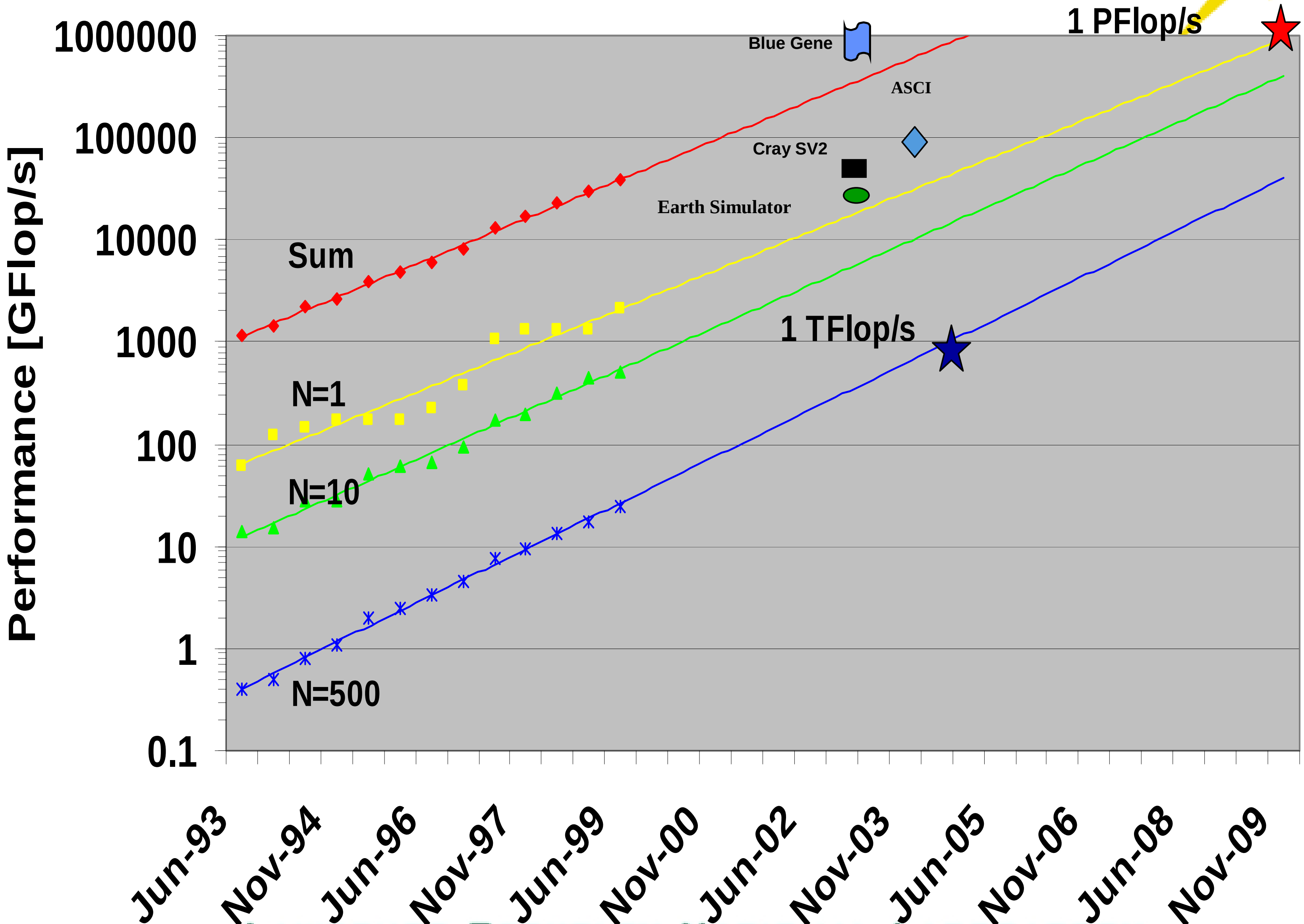


Analysis of TOP500 Data

- Annual performance growth about a factor of 1.82
- Two factors contribute almost equally to the annual total performance growth
- Processor number grows per year on the average by a factor of 1.30 and the
- Processor performance grows by 1.40 compared to 1.58 of Moore's Law

Strohmaier, Dongarra, Meuer, and Simon, Parallel Computing 25, 1999, pp 1517-1544.

Extrapolation to the Next Decade



Analysis of TOP500 Extrapolation

Based on the extrapolation from these fits we predict:

- **First 100~TFlop/s system by 2005**
- **About 1–2 years later than the ASCI path forward plans.**
- **No system smaller than 1 TFlop/s should be able to make the TOP500**
- **First Petaflop system available around 2009**
- **Rapid changes in the technologies used in HPC systems, therefore a projection for the architecture/technology is difficult**
- **Continue to expect rapid cycles of re-definition**

2001-2005: Technology Options

- Clusters
  - — SMP nodes, with custom interconnect
  - — PCs, with commodity interconnect
  - — vector nodes (in Japan)
- Custom built supercomputers
  - — Cray SV-2
  - — IBM Blue Gene
  - — HTMT
- Other technology options
  - — IRAM/PIM
  - — low power processors (Transmeta)
  - — consumer electronics (Playstation 2)
  - — Internet computing
  - — computational grids

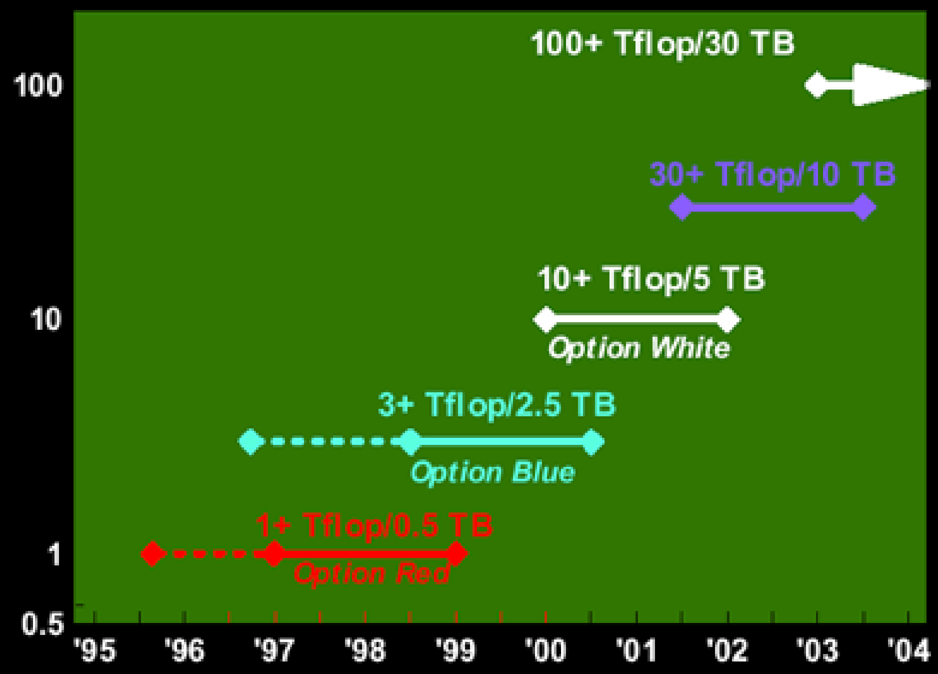
10 - 100 Tflop/s Cluster of SMPs

- The first ones are already on order
  - — LLNL installed a 10 Tflop/s in Sept. 2000
  - — NERSC installed a 3 Tflop/s system in Dec. 2000
  - — LANL will install a 30 Tflop/s Compaq system
- Systems are large clusters
  - — SMP nodes in US
  - — Vector nodes in Japan

- Programming model:
  - — OpenMP and/or vectors to maximize node speed
  - — MPI for global communication

Cluster of SMP Approach

- A Supercomputer is a "stretched" high-end server
  - parallel system, built by assembling nodes that are conventional, modest size, shared memory multiprocessor
  - just put more of them together

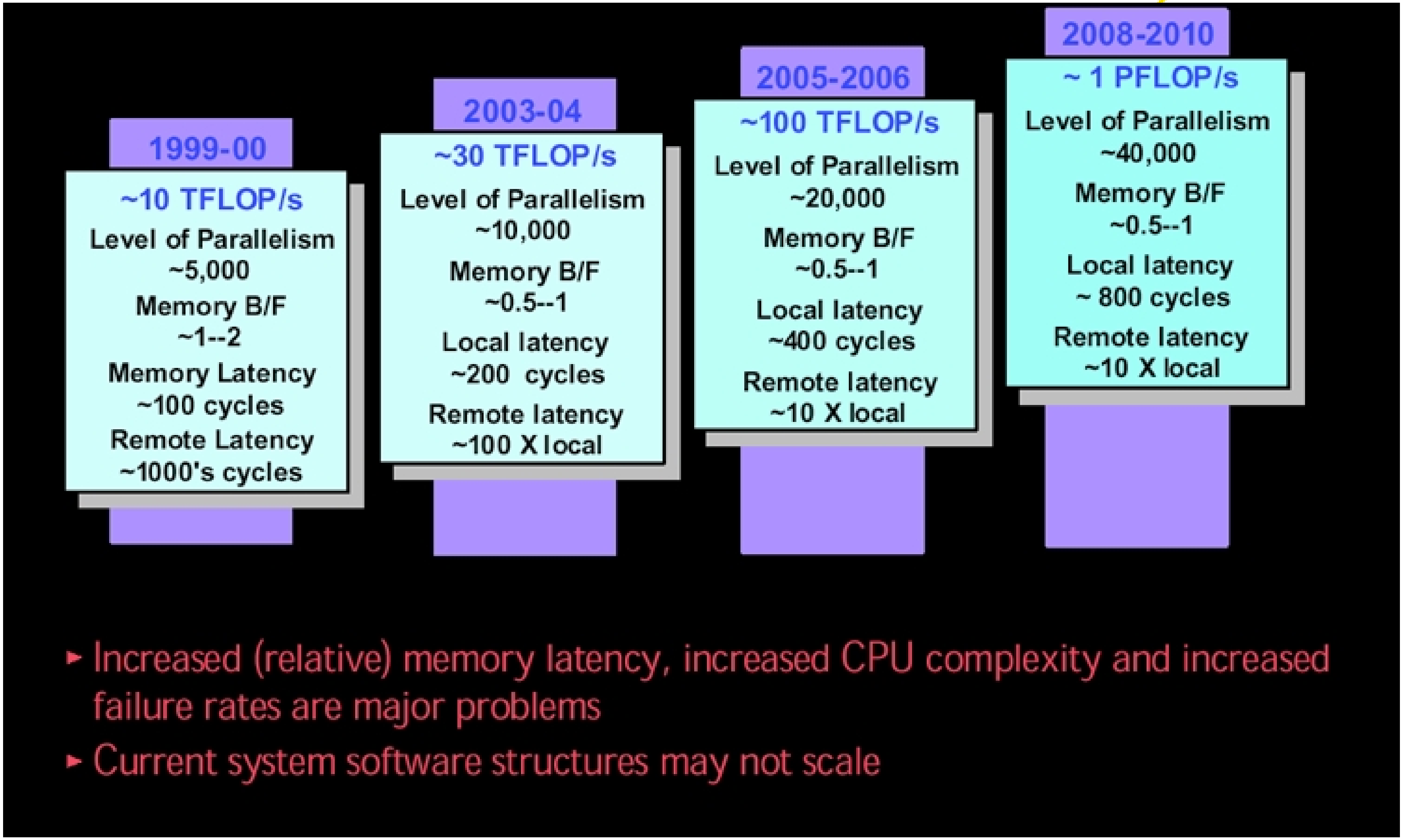


Accelerated Strategic Computing Initiative

ASCI Blue Pacific -- LLNL

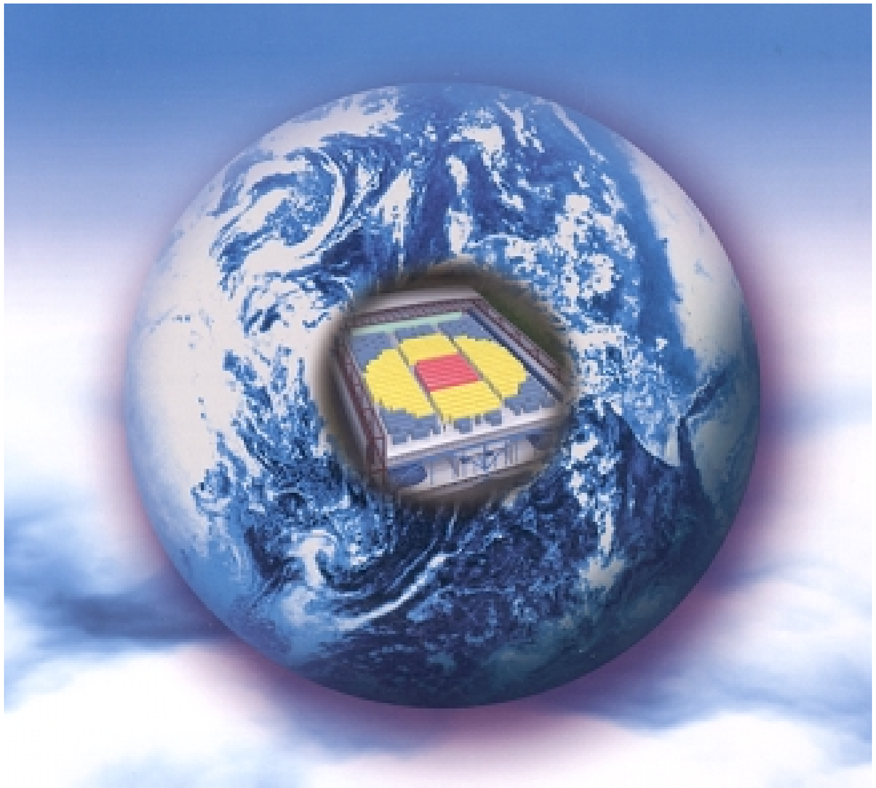
1,464 nodes; 5,856 CPUs
2.6 TB memory
80 TB disk
3.3 TFlop/s demonstrated

100 - 1000 Tflop/s Cluster of SMPs (IBM Roadmap)

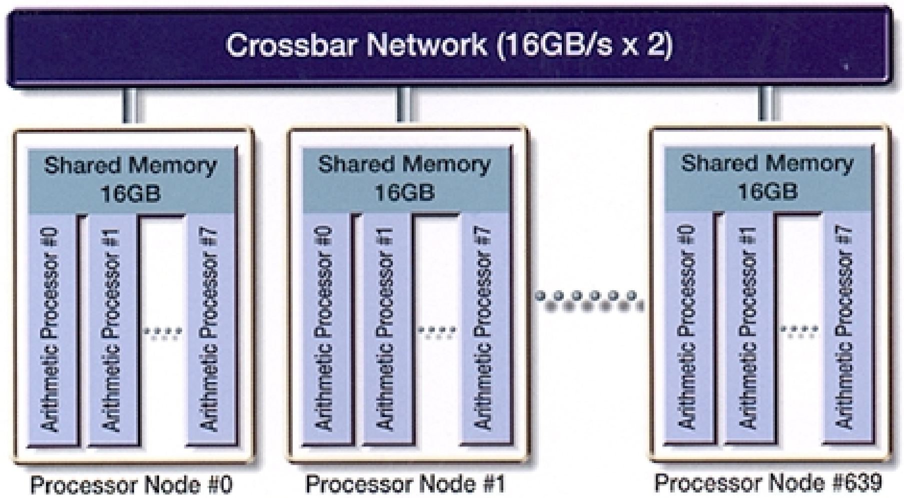


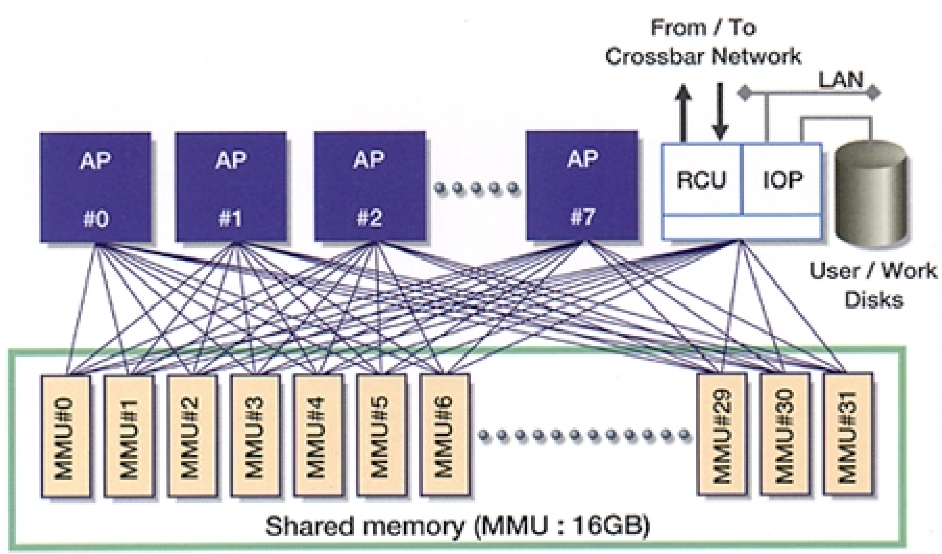
Earth Simulator

- 40 Tflop/s system in Japan
- completion 2002
- driven by climate and earthquake simulation requirements
- built by NEC
- 640 CMOS vector nodes



Earth Simulator





Cray SV2 Overview:

— Basic building block is a 50/100 GFLOPs node:

— 4 x CPUs per node. IEEE. Design goal is 12.8 GFLOPs per CPU.

— 8, 16 or 32 GB of coherent flat shared memory per CPU

— SSI to 1024 nodes: 50/100 TFLOPs, 32TB:

— 100 GB/sec interconnect capacity to/from each node

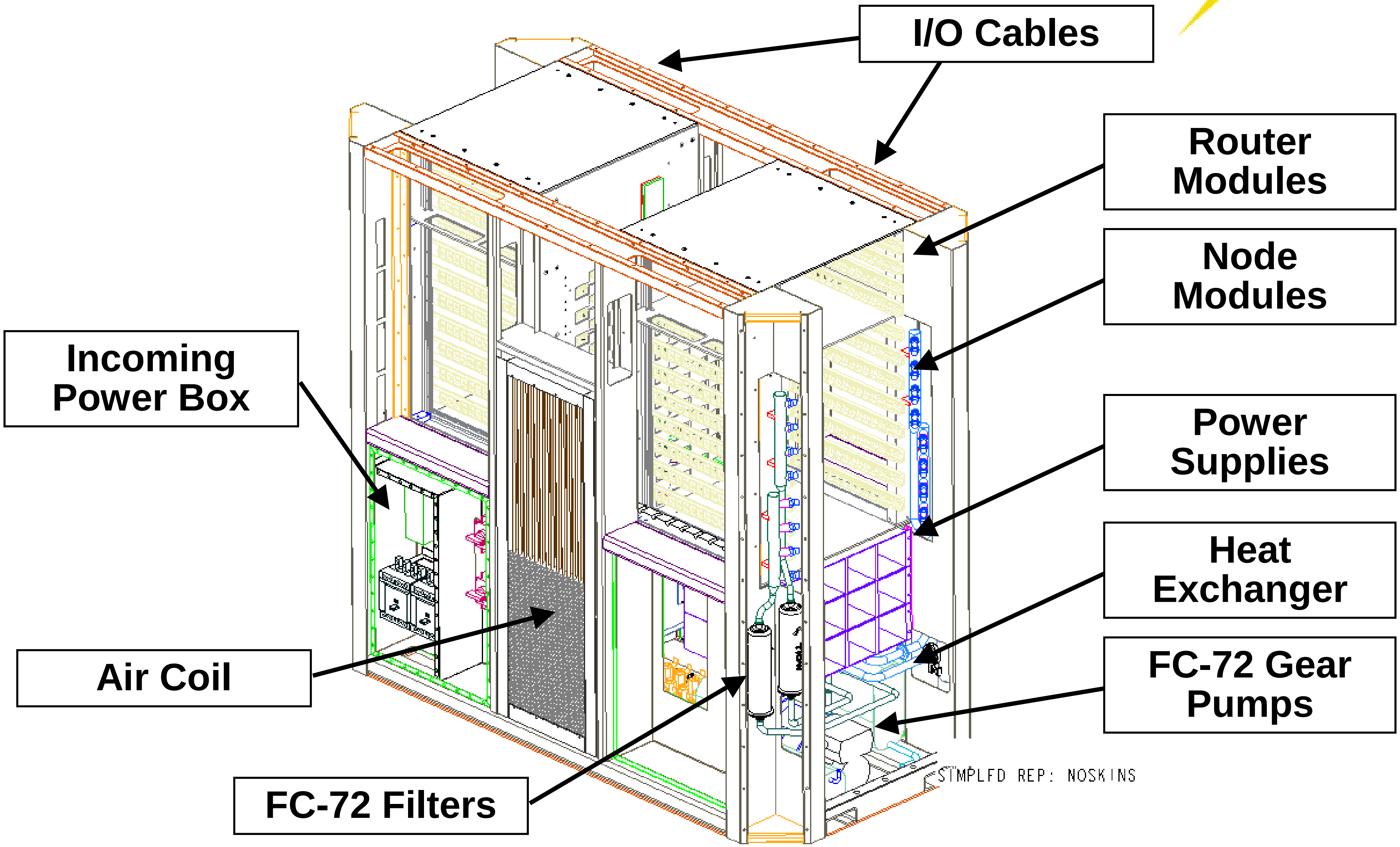
— ~1 microsecond latency anywhere in hypercube topology

— Targeted date of introduction, mid-2002.

— LC cabinets; Integral HEU (heat exchange unit)

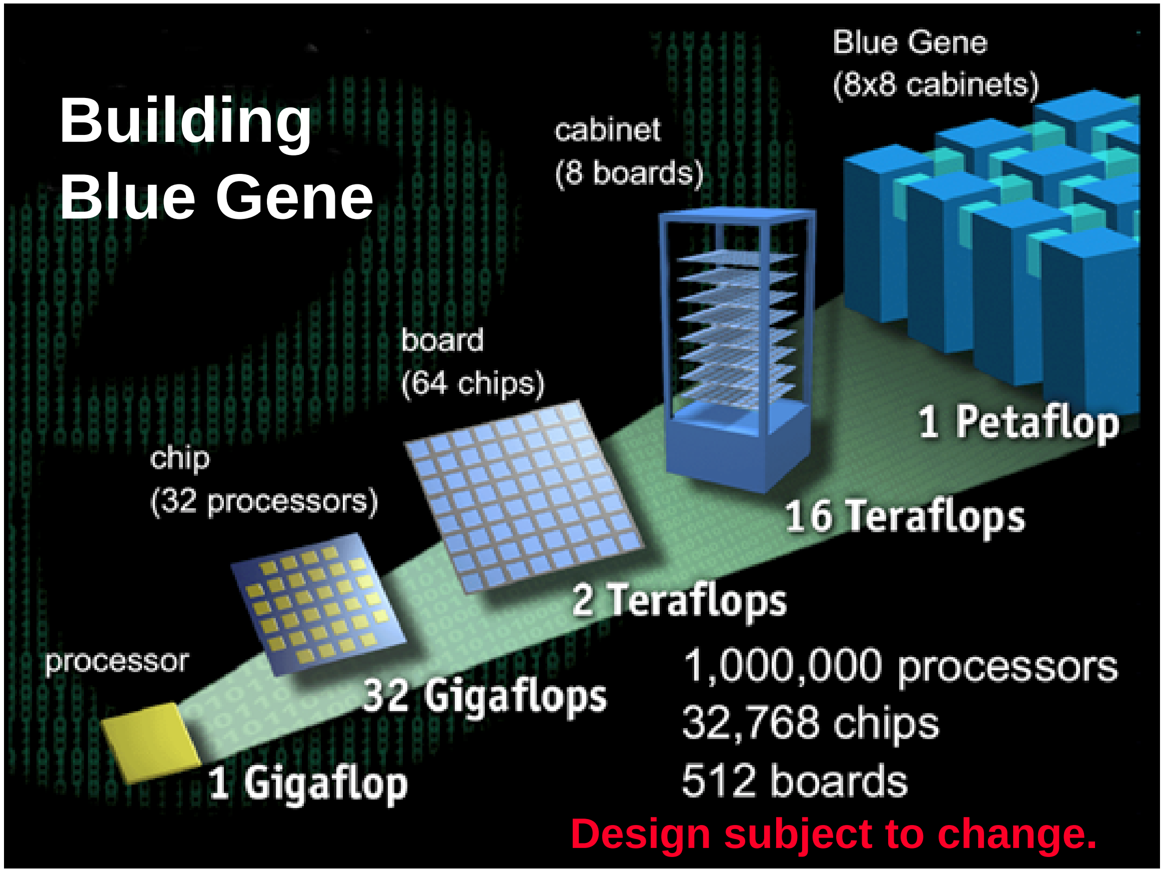
— Up to 64 cabinets (4096 CPUs/50 TFLOPS) mesh topology

Liquid-Cooled Cabinet — 64 CPUs



Cray Scalable Systems Update - Copyright Cray Inc, used by permission

CMOS Petaflop/s Solution



- IBM's Blue Gene
- 64,000 32 Gflop/s PIM chips
- Sustain O(10^7) ops/cycle to avoid Amdahl bottleneck

Five Computing Trends for the Next Five Years

- Continued rapid processor performance growth following Moore's law
- Open software model (Linux) will become standard
- Network bandwidth will grow at an even faster rate than Moore's Law
- Aggregation, centralization, colocation
- Commodity products everywhere

PC Clusters: Contributions of Beowulf

- An experiment in parallel computing systems
- Established vision of low cost, high end computing
- Demonstrated effectiveness of PC clusters for some (not all) classes of applications
- Provided networking software
- Conveyed findings to broad community (great PR)
- Tutorials and book
- Design standard to rally community!
- Standards beget: books, trained people, software … virtuous cycle

Adapted from Gordon Bell, presentation at Salishan

Linus's Law: Linux Everywhere

- Software is or should be free (Stallman)
- All source code is "open"
- Everyone is a tester
- Everything proceeds a lot faster when everyone works on one code (HPC: nothing gets done if resources are scattered)
- Anyone can support and market the code for any price
- Zero cost software attracts users!
- All the developers write lots of code
- Prevents community from losing HPC software (CM5, T3E)

Open Source Will Change the Rules!

- **Stage 1: (40s and 50s): every computer different, every program unique**
- **Stage 2: (60s and 70s): software is unbundled from hardware, commercial software companies arise**
- **Stage 3: (80s and 90s): mass market computers and mass market software, the notions of software copyright and privacy are born**
- **Stage 4: (2000 and beyond): software migrates to the WWW, OSS communities provide high quality software**

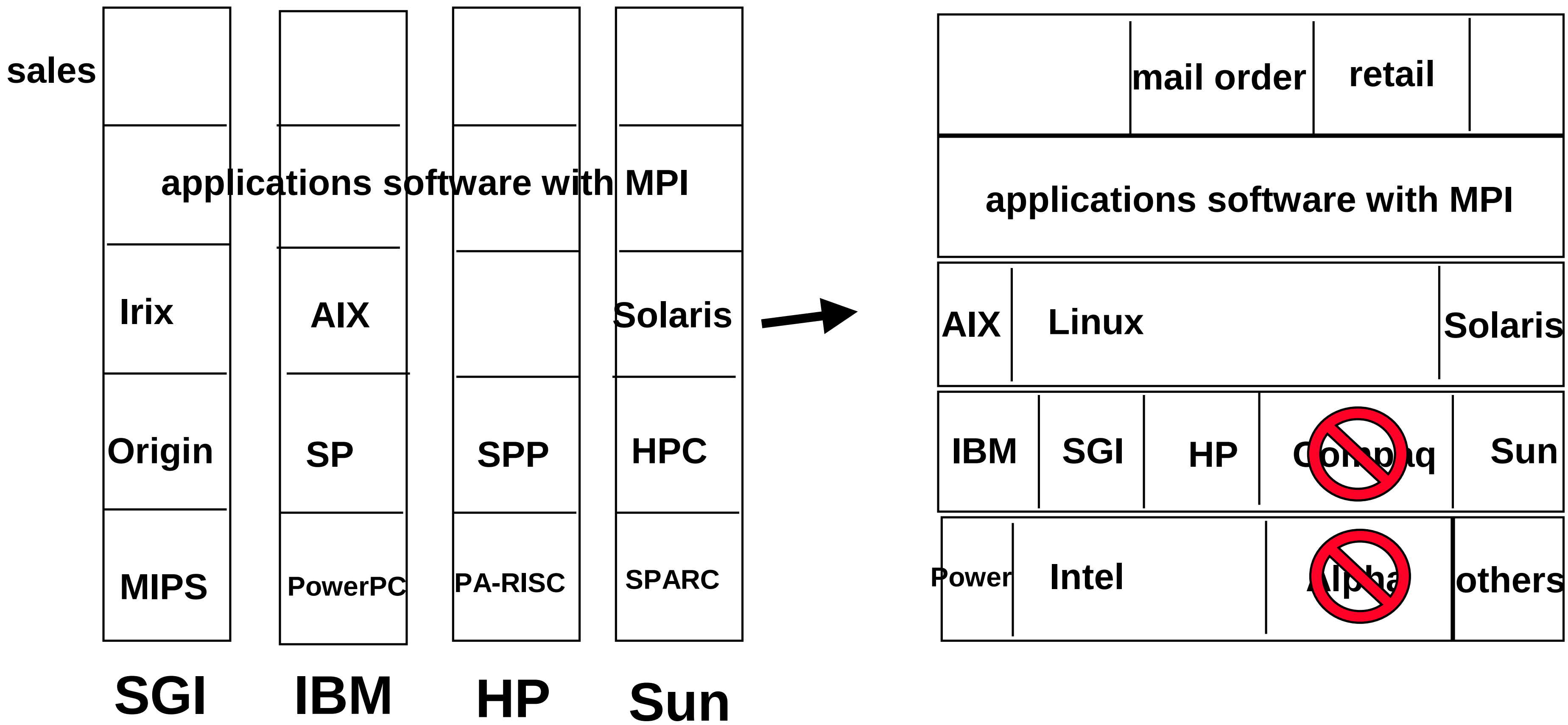
Commercially Integrated Clusters Are Already Happening

- **Forecast Systems Lab procurement (Prime contractor is High Performance Technologies Inc., subcontractor is Compaq)**
- **Los Lobos Cluster (IBM with University of New Mexico)**
- **NERSC has acquired a commercially integrated cluster in 2000 (IBM)**
- **Shell: largest engineering/scientific cluster**
- **NCSA: 1024 processor cluster (IA64)**
- **RWC Score Cluster**
- **DTF in US: 4 clusters for a total of 13 Teraflops (peak)**

2001-2005: Market Issues

From vertical to horizontal companies—

the Compaq-Dell model of High Performance Computing



Until 2010: A New Parallel Programming Methodology? - NOT

The software challenge: overcoming the MPI barrier

- **MPI created finally a standard for applications development in the HPC community**
- **Standards are always a barrier to further development**
- **The MPI standard is a least common denominator building on mid-80s technology**

Programming Model reflects hardware!

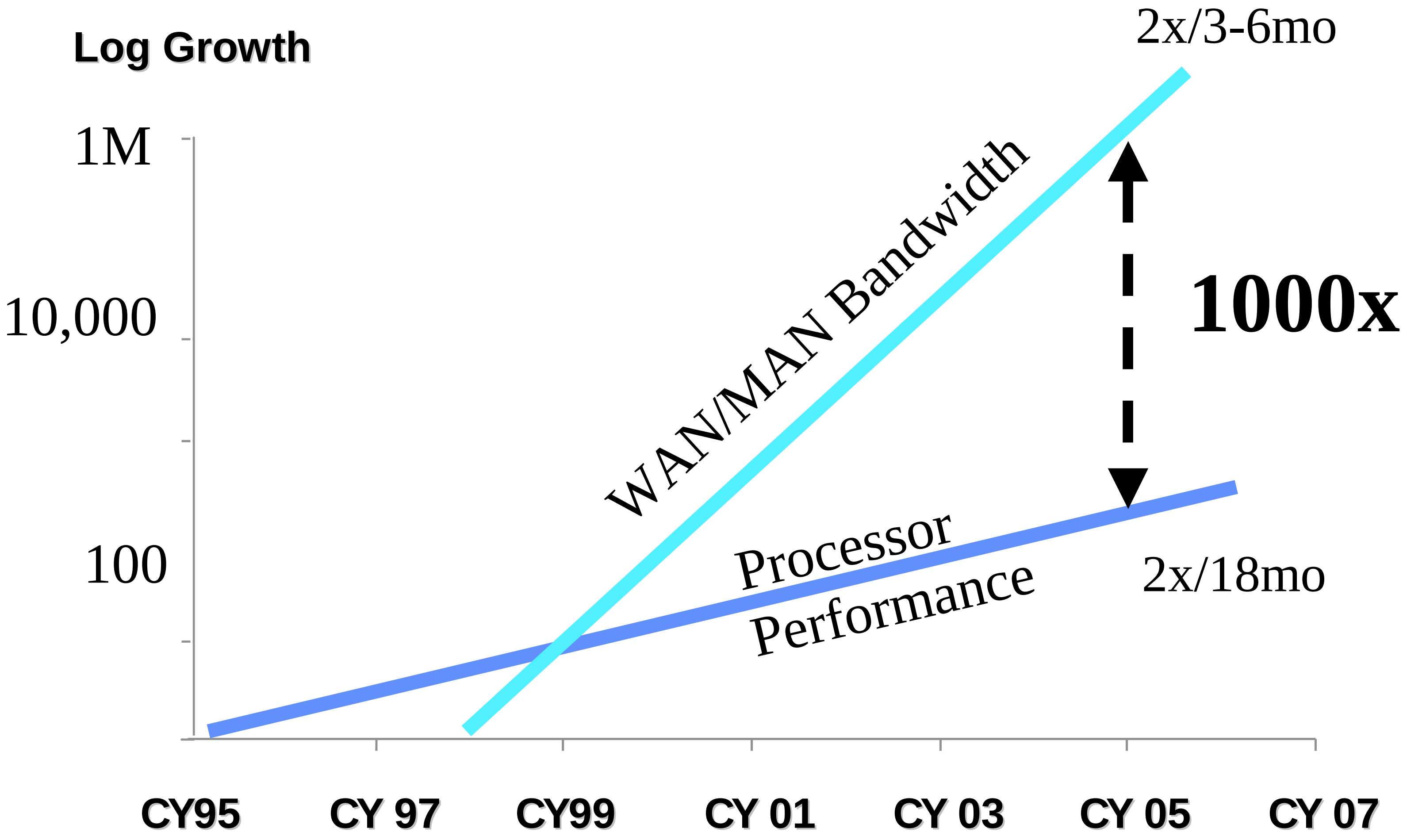
"I am not sure how I will program a Petaflops computer, but I am sure that I will need MPI somewhere" – HDS 2001

Five Computing Trends for the Next Five Years

- Continued rapid processor performance growth following Moore's law
- Open software model (Linux) will become standard
- Network bandwidth will grow at an even faster rate than Moore's Law
- Aggregation, centralization, colocation
- Commodity products everywhere

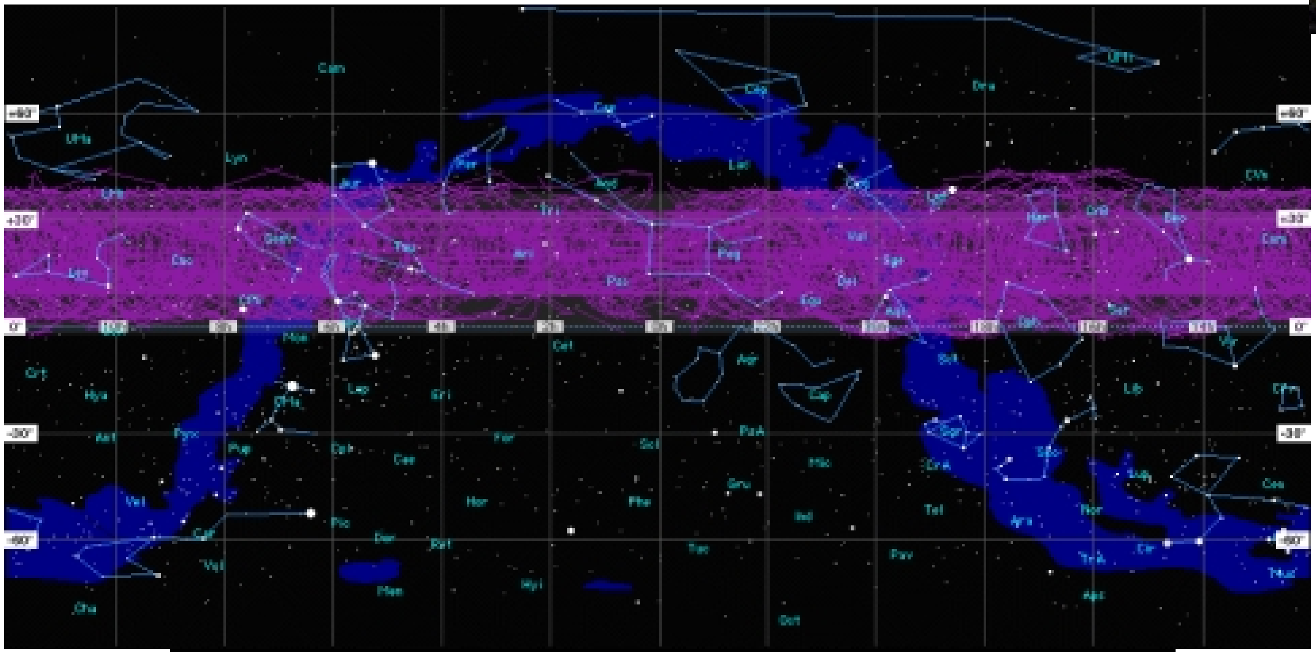
Bandwidth vs. Moore's Law

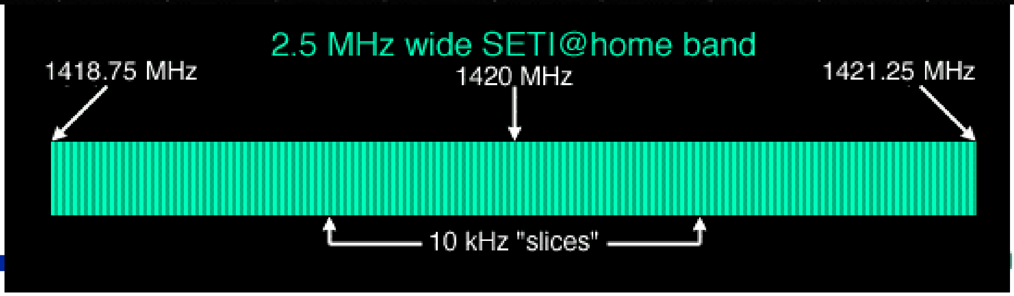
Adapted from G. Papadopoulos, Sun



Internet Computing- SETI@home

- Running on 500,000 PCs, ~1000 CPU Years per Day
  - — 485,821 CPU Years so far
- Sophisticated Data & Signal Processing Analysis
- Distributes Datasets from Arecibo Radio Telescope





Next Step-
Allen Telescope Array

The Vision for a DOE Science Grid

Scientific applications use workflow frameworks to coordinate resources and solve complex, multi-disciplinary problems

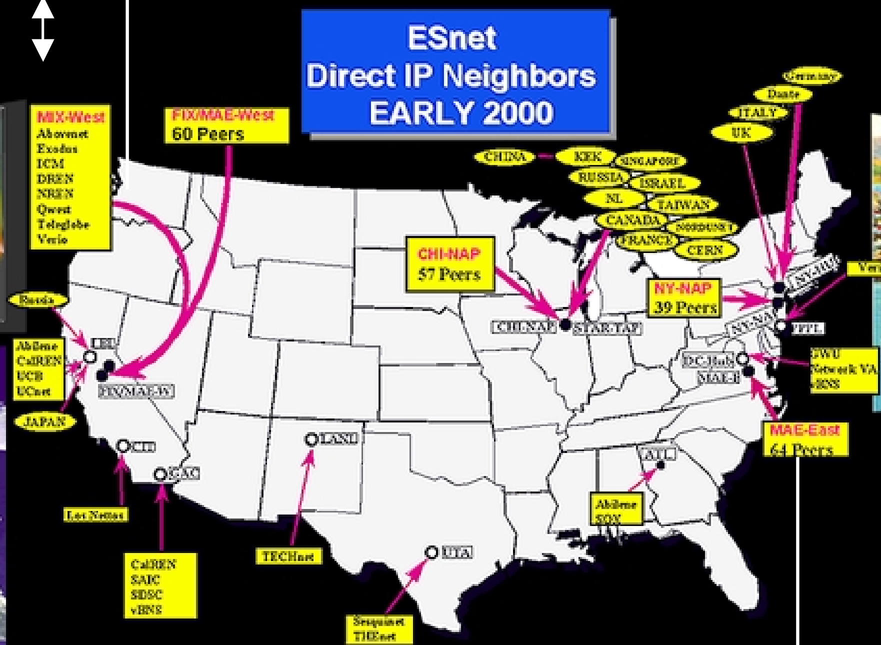
Grid services provide a uniform view of many diverse resources

Large-scale science and engineering is typically done through the interaction of

- People,
- Heterogeneous computing resources,
- Multiple information systems, and
- Instruments

All of which are geographically and organizationally dispersed.

The overall motivation for “Grids” is to enable the routine interactions of these resources to facilitate this type of large-scale science and engineering.



Two Sets of Goals

Our overall goal is to facilitate the establishment of a DOE Science Grid (“DSG”) that ultimately incorporates production resources and involves most, if not all, of the DOE Labs and their partners.

A “local” goal is to use the Grid framework to motivate the R&D agenda of the LBNL Computing Sciences, Distributed Systems Department (“DSD”).

Impact on HPC

- Internet Computing will stay on the fringe of HPC
  — no viable model to make it commercially realizable

- Grid activities will provide an integration of data, computing, and experimental resources
  — but not metacomputing

- More bandwidth will lead to aggregation of HPC resources, not to distribution

Five Computing Trends for the Next Five Years

- Continued rapid processor performance growth following Moore’s law
- Open software model (Linux) will become standard
- Network bandwidth will grow at an even faster rate than Moore’s Law
- Aggregation, centralization, co-location
- Commodity products everywhere

A “Supercomputing” Center in 2006

http://sanjose.bcentral.com/sanjose/stories/2001/03/19/daily51.html

March 22, 2001

Huge server farm proposed for San Jose

What is being billed as the largest server farm in the world is heading for city approval in San Jose. If built as planned on a campus in the Alviso area of the city, the server farm would use 150 megawatts of power from the state's power grid plus 30 megawatts generated on site.

But officials of Pacific Gas and Electric Co. say they cannot supply the needed power at this time.

The server farm proposed by U.S. DataPort of San Jose would cost about $1.2 billion to construct, encompassing 10 buildings on a 170 acre campus and would handle as much as 15 percent of the world's entire Internet traffic. It would take about five years to build out -- enough time company officials hope, for the state to solve the current electricity shortages.

Server farms are concentrations of computers and related equipment which handle Internet-related chores. In addition to needing power for the computers, telephone switches, routers and other equipment, they need power for air conditioning to cool the buildings.

The city planning commission has given its preliminary approval to the plans. Final action is expected in April.

NERSC's Strategy Until 2010: Oakland Scientific Facility

New Machine Room — 20,000 ft^{2,} Option open to expand to 40,000 ft$^{2.}$

Includes ~50 offices and 6 megawatt electrical supply.

It's a deal: \$1.40/ft^2 when Oakland rents are >\$2.50/ ft^2 and rising!

The Oakland Facility Machine Room

Power and cooling are major costs of ownership of modern supercomputers

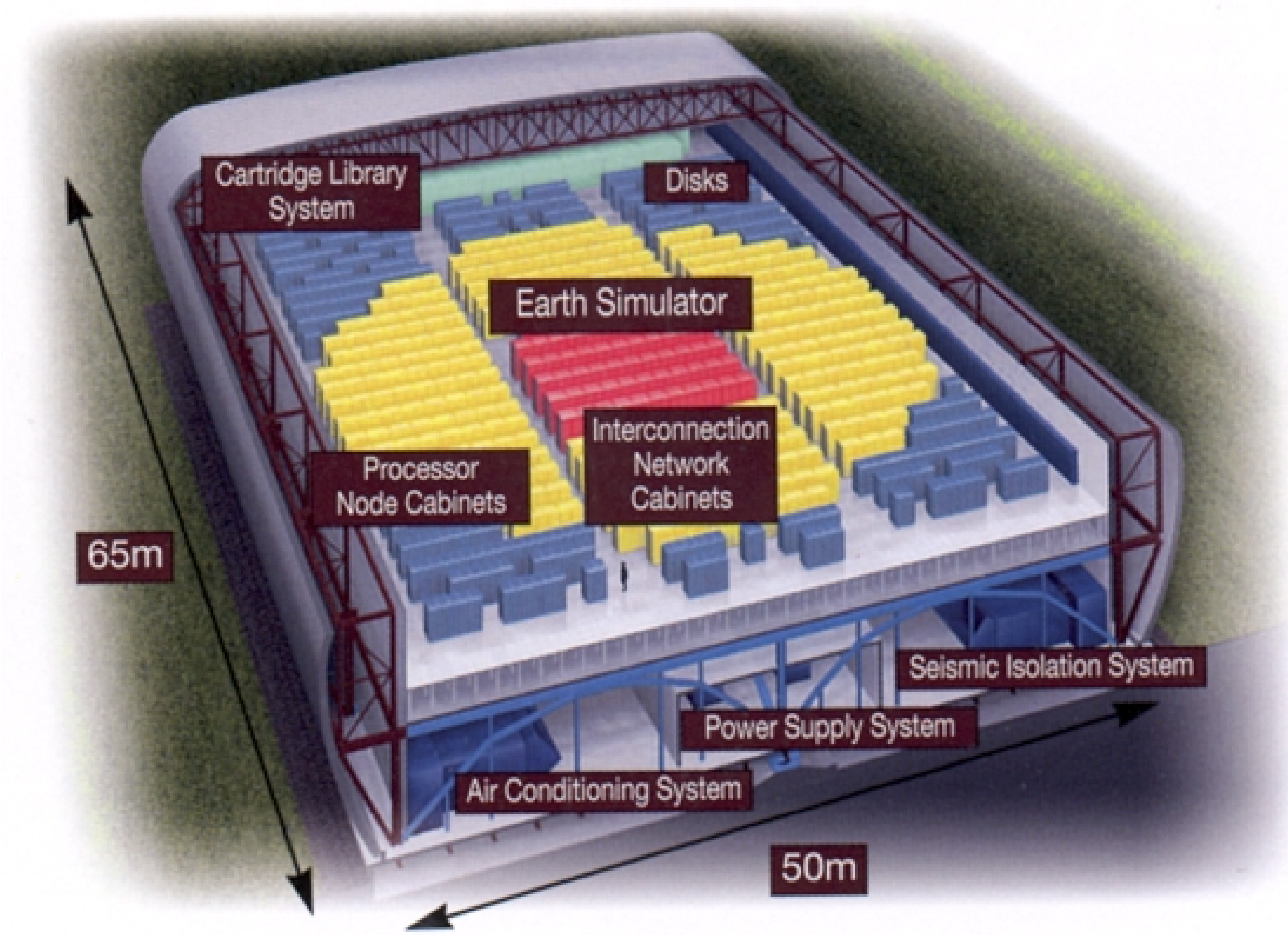
Expandable to 6 Megawatts

Strategic Computing Complex at LANL – home of the 30 Tflop/s Q machine

Strategic Computing Complex at LANL

- 303,000 gross sq. ft.
- 43,500 sq. ft. unobstructed computer room
  - — Q consumes approximately half of this space
- 1 Powerwall Theater (6X4 stereo = 24 screens)
- 4 Collaboration rooms (3X2 stereo = 6 screens)
  - — 2 secure, 2 open (1 of each initially)
- 2 Immersive Rooms
- Design Simulation Laboratories (200 classified, 100 unclassified)
- 200 seat auditorium

Earth Simulator Building



“I used to think computer architecture was about how to organize gates and chips – not about building computer rooms”

Thomas Sterling, Salishan, 2001

For the Next Decade, The Most Powerful Supercomputers Will Increase in Size

This

Became

And will get bigger

Power and cooling are also increasingly problematic, but there are limiting forces in those areas.

- Increased power density and RF leakage power, will limit clock frequency and amount of logic *[Shekhar Borkar, Intel]*
- So linear extrapolation of operating temperatures to Rocket Nozzle values by 2010 is likely to be wrong.

Five Computing Trends for the Next Five Years

- Continued rapid processor performance growth following Moore's law
- Open software model (Linux) will become standard
- Network bandwidth will grow at an even faster rate than Moore's Law
- Aggregation, centralization, co-location
- Commodity products everywhere

…. the first ever coffee machine to send e-mails

"Lavazza and eDevice present the first ever coffee machine to send e-mails

On-board Internet connectivity leaves the laboratories

eDevice, a Franco-American start-up that specializes in the development of on-board Internet technology, presents a world premiere: e-espressopoint, the first coffee machine connected directly to the Internet. The project is the result of close collaboration with Lavazza, a world leader in the espresso market with over 40 million cups drunk each day.

Lavazza's e-espressopoint is a coffee machine capable of sending e-mails in order, for example, to trigger maintenance checks or restocking visits. It can also receive e-mails from any PC in the given service.

A partnership bringing together new technologies and a traditional profession …"

See http://www.cyperus.fr/2000/11/edevice/cpuk.htm

New Economic Driver: IP on Everything

Source: Gordon Bell, Microsoft, Lecture at Salishan Conf.

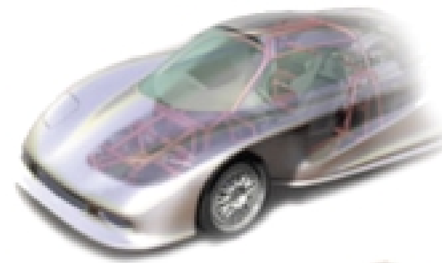
Enablers of Pervasive Technologies

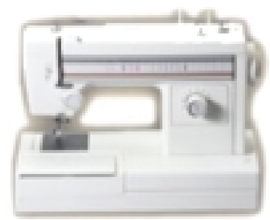
- General accessibility through intuitive interfaces
- A supporting infrastructure, perceived valuable, based on enduring standards
- MOSAIC browser and World Wide Web are enablers of global information infrastructure

Source: Joel Birnbaum, HP, Lecture at APS Centennial, Atlanta, 1999

Information Appliances

- Are characterized by what they do
- Hide their own complexity
- Conform to a mental model of usage
- Are consistent and predictable
- Can be tailored
- Need not be portable



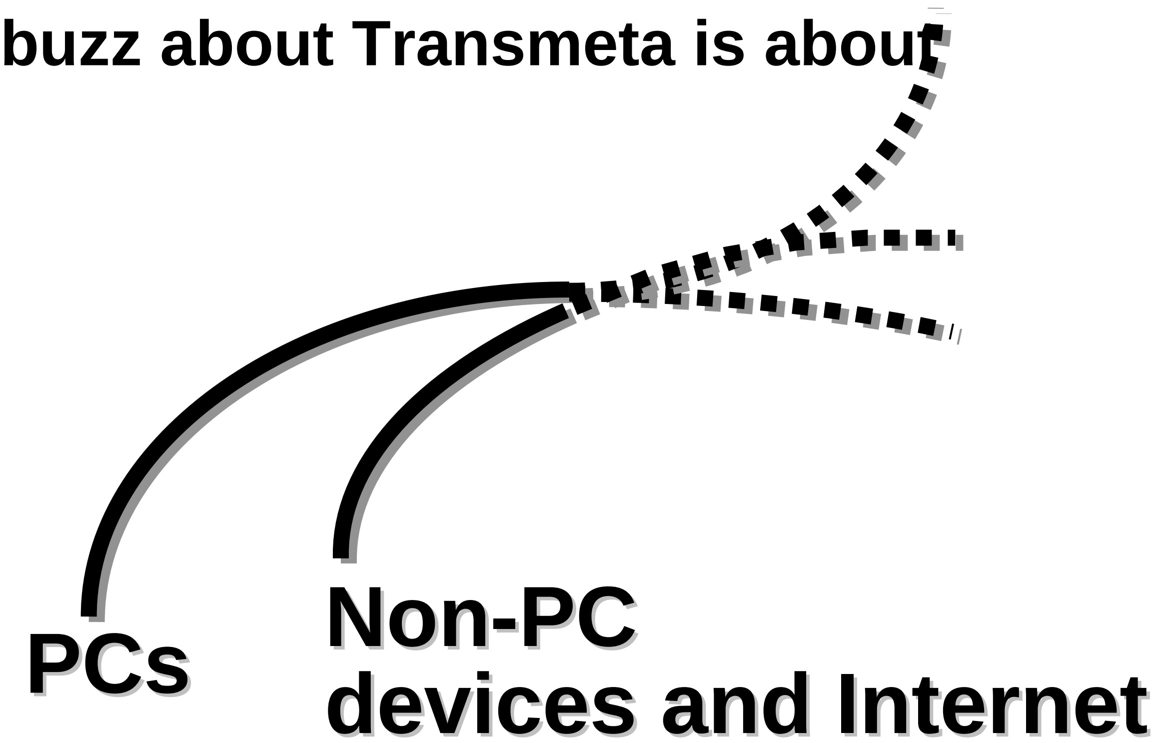


Source: Joel Birnbaum, HP, Lecture at APS Centennial, Atlanta, 1999

… but what does that have to do with supercomputing?

HPC depends on the economic driver from below:

- Mass produced cheap processors will bring microprocessor companies increased revenue
- system on a chip will happen soon
- that is what the buzz about Transmeta is about



"PCs at Inflection Point", Gordon Bell, 2000

ISTORE Hardware Vision

System-on-a-chip enables computer, memory, without significantly increasing size of disk

5-7 year target:

MicroDrive:1.7" x 1.4" x 0.2"
2006: ?
1999: 340 MB, 5400 RPM,
5 MB/s, 15 ms seek
2006: 9 GB, 50 MB/s ?   (1.6X/yr capacity, 1.4X/yr BW)

Integrated IRAM processor
2x height

Connected via crossbar switch
growing like Moore's law

16 Mbytes; ; 1.6 Gflops; 6.4 Gops

10,000+ nodes in one rack! 100/board = 1 TB; 0.16 Tf

Source: David Patterson, UC Berkeley

What am I willing to predict?

2010:

- **Petaflop (peak) supercomputer before 2010**
- **We will use MPI on it**
- **It will be built from commodity parts**
- **I can't make a prediction from which technology (systems on a chip to "SMP servers" are possible)**
- **The "grid" will have happened, because a killer app made it commercially viable**
- **An incredible tale like:**
  - **Microsoft will be split into three companies; in 2005 the Microsoft applications company buys Cray Inc.; $$ are spent in revamping the Tera MTA; the company loses focus on its key applications; word processing, spreadsheets etc. are provided by open source competitors ...**
- **Disruption of all this because of unrelated outside development, for example a boom in robotics starting in 2005**